علیرضا آخوندی

9731107

سوال اول :

Dimension : بعد به معنای ویژگی های یک شی دیتایی در دیتاست می باشد و در واقع می توان گفت هم ارز فیچرهای هر شی است.

Outlier : به داده‌های گفته می‌شود که به مانند نویز نیستند اما رفتاری متفاوت از سایر داده‌های موجود در توزیع دارند. این داده ها کمتر از lower bound و بیشتر از upper bound هستند و داده پرت محسوب می‌شوند. رابطه lower bound و upper bound به صورت زیر است.

$$lower\ bound = Q_1 - 1.5 \times IQR$$
$$upper\ bound = Q_3 + 1.5 \times IQR$$

Independent Variable : به ورودی های یک فرآیند تجزیه و تحلیل بر روی آنها انجام می شود و ما دنبال پیدا کردن رابطه بین آن ها هستیم.

Dependent Variable: به خروجی فرایند تجزیه و تحلیل روی متغیر های مستقل گویند.

Stratified Sampling : به نوعی نمونه برداری گفته می‌شود که پس از کلاس بندی کردن و قرار دادن آن ها در گروه های مختلف، اقدام به نمونه برداری رندوم از هر گروه بکنیم. به این صورت هر گروه یک زیر جامعه از جامعه اصلی می شود.

سوال دوم:

Decimal Scaling: در این روش ابتدا عددی که بزرگتری قدر مطلق را دارد را پیدا کرده و تعداد ارقام آن را j می نامیم. حال تمامی اعداد در جمعیت را تقسیم بر 10^j می کنیم. این کار اعداد را بین -1 تا 1 مپ میکند.

Zero Score: در این روش انحراف معیار و میانگین جمعیت را به دست می آوریم. سپس با کمک فرمول زیر مقادیر را بروز می کنیم.

$$x' = \frac{x - \mu}{\sigma}$$

Min-max : در این روش مقادیر جمعیت را با کمک مینیمم و ماکسیمم بین صفر تا یک نگاشت میکنیم.

$$x' = \frac{x - \min(X)}{\max(X) - \min(X)}$$

سوال سوم :

فرایند این الگوریتم به این صورت است که در ابتدا به صورت دلخواه یک بازه بندی انتخاب می شود و آن را به بازه های با اندازه کوچک تر به طول 2 یا 3 یا 4 یا ... تبدیل میکنیم. پس از آن، با فرمولی که در الگوریتم عنوان شده مقدار شباهت این بازه ها را نسبت به یکدیگر اندازه گیری میکنیم و خروجی آن فرمول را

Chi-square می نامیم. پس از مقایسه خروجی ها، دو خروجی که نزدیک ترین عدد به هم هستند را با هم ادغام می کنیم و دوباره الگوریتم را تکرار می کنیم تا شرط خاتمه برقرار شود به این معنا که همه بازه ها به خروجی Chi-square یکسان برسند.

سوال چهارم :

سوال چهارم:

الف)

$x = [1,1,1,1]$, $y = [2,2,2,2]$

Cosine: $x \cdot y = 8$ $\quad$ $Cosine(x,y) = \frac{x \cdot y}{\|x\| \|y\|} = \frac{8}{2 \times 4} = 1$

$Corr(x,y) = \frac{Covariance(x,y)}{standard\_dev(x) \times standard\_dev(y)}$

$std(x) = std(y) = 0$ ⟵ بی‌قابل تعریف

$x = [0,1,0,1]$ $\quad$ $y = [1,0,1,0]$

$Cosine(x,y) = 0$ $\quad$ $euclidean(x,y) = \sqrt{(2-1)^2 + (2-1)^2 + (2-1)^2 + (2-1)^2} = 2$

ب)

$x = [0,1,0,1]$ $\quad$ $y = [1,0,1,0]$

$Cosine(x,y) = 0$

correlation: $Corr(x,y) = \frac{S_{xy}}{S_x S_y}$

$S_x = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2} = \sqrt{\frac{1}{3} \sum_{k=1}^{4} (x_k - 0.5)^2}$

$= \sqrt{\frac{1}{3} \times 1} = \frac{1}{\sqrt{3}}$

$S_y = \frac{1}{\sqrt{3}}$ $\quad$ $S_{xy} = \frac{1}{n-1} \sum_{k=1}^{4} (x_k - \bar{x})(y_k - \bar{y})$

$= \frac{1}{3} \times (-0.5 \times 0.5 \times 4)$

$Corr(x,y) = -1$

(1

Euclidean$(x,y) = \sqrt{1+1+1+1} = 2$

Jaccard $= \frac{|A \cap B|}{|A \cup B|} = 0$

(2

$x = [1,1,0,1,0,1]$, $y = [1,1,1,0,0,1]$

Covariance$(x,y) = \frac{1}{5}\sum_{i=1}^{6}(x_i - \frac{2}{3})(y_i - \frac{2}{3})$

$= \frac{1}{5}\left(\frac{1}{9} + \frac{1}{9} - \frac{2}{9} - \frac{2}{9} + \frac{4}{9} + \frac{1}{9}\right) = \frac{1}{15}$

$S_x = \sqrt{\frac{1}{5}\sum_{k=1}^{6}(x_i - \bar{x})^2} = \sqrt{\frac{1}{5}\left(\frac{1}{9} + \frac{1}{9} + \frac{4}{9} + \frac{4}{9} + \frac{1}{9} + \frac{1}{9}\right)}$

$= \frac{2}{\sqrt{15}}$

$Corr(x,y) = \frac{\frac{1}{15}}{\frac{4}{15}} = \frac{1}{4}$

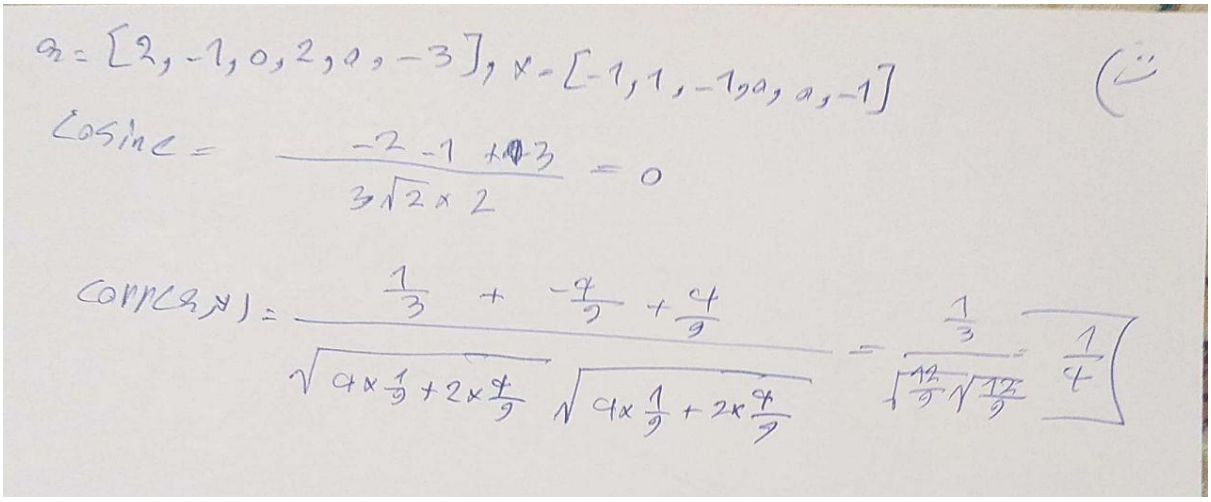
manhattan :

$d(x,y) = |1-1| + |1-1| + |0-1| + |1-0| + |0-0| + |1-1| = 2$

Bhattacharyya distance :

$BC(x,y) = \sqrt{1\times1} + \sqrt{1\times1} + \sqrt{0\times1} + \sqrt{1\times0} + \sqrt{0\times0} + \sqrt{1\times1} = 3$

$d_B(x,y) = -\ln(BC(x,y)) = -\ln 3$



سوال پنجم :

اگر مجموعه داده ها بسیار بزرگ باشد، تجزیه و تحلیل پیچیده داده ها و استخراج حجم عظیمی از داده ها می‌تواند زمان بسیار زیادی ببرد. لذا باید از تکنیک کاهش داده استفاده کنیم. هدف تکنیک کاهش داده به دست آوردن یک نمایش کوچک شده از دیتاست اصلی ما می باشد، که این دیتاست جدید از دیتاست اصلی حجم کمتری دارد اما این دیتاست جدید یکپارچگی داده های اصلی را نیز حفظ کرده است. پس استخراج بر روی دیتاست های کاهش یافته باید کارآمدتر باشد و در عین حال نتایج تحلیلی تقریبا یکسان مانند دیتاست اصلی ایجاد کند. استراتژی و راهبردهای کاهش داده شامل کاهش تعداد، کاهش ابعاد و فشرده سازی داده ها می باشد.

کاهش ابعاد فرآیند کاهش تعداد متغیرها یا ویژگی های تصادفی مورد بررسی است. روش های کاهش ابعاد شامل :

- تبدیل موجک (wavelet transform)
- تجزیه و تحلیل مولفه های اصلی **(PCA)**
- انتخاب زیرمجموعه ویژگی و ایجاد ویژگی

می باشد. روش های کاهش تعداد از مدل‌های پارامتریک یا ناپارامتریک برای به دست آوردن نمایش‌های کوچک‌تر از داده‌های اصلی استفاده می‌کنند. مدل های پارامتریک به جای داده های واقعی، فقط پارامترهای مدل را ذخیره می کنند. به عنوان مثال رگرسیون از این نوع می باشند. همچنین روش‌های غیرپارامتریک شامل هیستوگرام، خوشه‌بندی، نمونه‌برداری و تجمع مکعب داده‌ها است. روش‌های فشرده‌سازی داده‌ها، تبدیل‌ها را برای به دست آوردن نمایشی کاهش‌یافته یا فشرده از داده‌های اصلی اعمال می‌کنند. در این روش ها اگر بتوان داده های اصلی را بدون از دست دادن

اطلاعات از روی داده های فشرده بازسازی کرد، کاهش داده بدون از دست دادن (lossless) است، در غیر این صورت زیان آور (lossy) است.

سوال ششم: در این روش یک بردار را به یک بردار عددی متفاوت با همان طول و اندازه از ضرایب موجک تبدیل می کنیم.

تفاوت feature selection و feature extraction :

دلیل استفاده از feature selection، فیلتر کردن ویژگی های نامربوط یا اضافی از دیتاست اصلی است. تفاوت اصلی بین feature selection و feature extraction در این است که feature selection زیر مجموعه ای از ویژگی های اصلی را حفظ میکند و از طرفی دیگر feature extraction ویژگی های جدید ایجاد می کند.

سوال هفتم :

در ابتدا داده هارا به شکل صعودی مرتب می کنیم :

20, 44, 56, 70, 71, 73, 74, 74, 80, 80, 89, 89, 90, 90, 100, 143, 143, 144, 146

با توجه به داده ها می توان دید که :

Min = 20, Max = 146, Median = 80

همچنین :

$$Q_1 = 71\,, Q_3 = 100\,, IQR = 100 - 71 = 19$$

همچنین میدانیم که :

$$lower\ bound = Q_1 - 1.5\ \times\ IQR = 27.5$$
$$upper\ bound = Q_3 + 1.5\ \times IQR = 143.5$$

در اینجا داده های 20 – 144 – 146 داده های پرت هستند.

در این حالت پس از حذف این داده ها مینیمم و ماکسیمم به ترتیب 44 و 143 می شوند.

سوال هشتم :

1. خیر، زیرا نویز در ویژگی ها به طور پیش فرض نامطلوب می باشد، چرا که مقادیر ویژگی ها را تحریف میکنند اما در خصوص outlier می توان گفت در مواقعی مطلوب هستند، چرا که در برخی از تسک ها هدف ما شناسایی آنها است.
2. بله، زیرا نوز می تواند داده ها را تصادفی تر بکند، بنابراین ممکن است برخی از نمونه ها در یک مجموعه داده نویزی به صورت outlier ظاهر شوند. به عنوان مثال اگر مقدار داده نویزی از مقادیر مجموعه داده عمومی فاصله بیش از حدی داشته باشد، این داده هم نویز و هم outlier محسوب می شود.
3. خیر، زیرا داده های پرت می توانند مقادیر واقعی باشند که به نظر می رسد که به دیتاست ما تعلق ندارند.
4. بله، هنگامی که نویز رخ می دهد مقدار یک ویژگی تحریف می شود، حال مقدار این ویژگی می تواند مقدار عادی باشد و در اثر نویز به oulier تبدیل شود یا بالعکس.

سوال نهم:

الف) در صورت مثبت بودن ورودی ها، بازه صفر تا یک در غیر این صورت بازه منفی یک تا مثبت یک.

ب) خیر، زیرا میتوان نتیجه گرفت که دو بردار هم راستا و هم جهت هستند و زاویه بین آن ها صفر است.

ج) برای دو بردار x و y اگر میانگین هر دو بردار برابر صفر باشد داریم :

$$Correlation(x,y) = \frac{\sum_{i=1}^{n}(x_i - 0)(y_i - 0)}{\sqrt{\sum_{i=1}^{n}(x_i - 0)^2}\sqrt{\sum_{i=1}^{n}(y_i - 0)^2}} = \frac{x.y}{\|x\|\,\|y\|} = cosine\ similarity(x,y)$$

سوال دهم:

نمودار quantile برای نشان دادن درصد تقریبی مقادیر کمتر یا برابر با متغیر مستقل در توزیع تک متغیره استفاده می شود. بنابراین، اطلاعات quantile را برای همه داده‌ها نمایش می‌دهد، جایی که مقادیر اندازه‌گیری شده برای متغیر مستقل در برابر quantile متناظر آنها رسم می‌شود. در حالی که نمودار quantile-quantile کوانتایل های یک توزیع تک متغیره را در برابر کوانتایل های متناظر یک توزیع تک متغیره دیگر نمودار می کند.

سوال یازدهم:

برای ویژگی numeric می توان از فاصله اقلیدسی، فاصله منهتن یا فاصله supremum استفاده کرد. اگر x و y را دو بردارد در نظر بگیریم که هرکدام n مولفه داشته باشند، فاصله اقلیدسی به صورت زیر تعریف می شود:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

همچنین برای محاسبه فاصله منهتن نیز از بردار زیر استفاده می کنیم :

$$d(x,y) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$

ویژگی های nominal میتوانند دو یا بیشتر مقدار متفاوت داشته باشند. برای محاسبه عدم تشابه بین اشیا توصیف شده توسط ویژگی nominal می توان بر اساس فرمول زیر عمل کرد که در آن m برابر با تعداد مج ها و p تعداد کل متغیر هاست.

$$d(x,y) = \frac{p - m}{p}$$